



**UNIVERSIDADE DE BRASÍLIA**  
**INSTITUTO DE LETRAS**  
**DEPARTAMENTO DE LÍNGUAS ESTRANGEIRAS E TRADUÇÃO**  
**LÍNGUAS ESTRANGEIRAS APLICADAS AO MULTILINGUISMO E**  
**À SOCIEDADE DA INFORMAÇÃO**

**GUSTAVO ROBERTO DA SILVA GOMES**

***AMPLIAÇÃO DO GLOSSÁRIO MULTILÍNGUE ONLINE SOBRE MIGRAÇÃO***  
***E REFÚGIO: O PROCEDIMENTO DE CRIAÇÃO DE NOVAS FICHAS***

Brasília  
Dezembro de 2020

GUSTAVO ROBERTO DA SILVA GOMES

**AMPLIAÇÃO DO *GLOSSÁRIO MULTILÍNGUE ONLINE SOBRE MIGRAÇÃO*  
*E REFÚGIO*: O PROCEDIMENTO DE CRIAÇÃO DE NOVAS FICHAS**

Trabalho de Conclusão de Curso apresentado ao Departamento de Línguas Estrangeiras e Tradução como requisito parcial para a obtenção do título de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI).

Orientadora: Prof.<sup>a</sup> Dr.<sup>a</sup> Elisa Duarte Teixeira.

BRASÍLIA, DF

2020

GUSTAVO ROBERTO DA SILVA GOMES

AMPLIAÇÃO DO GLOSSÁRIO MULTILÍNGUE ONLINE SOBRE MIGRAÇÃO  
E REFÚGIO: O PROCEDIMENTO DE CRIAÇÃO DE NOVAS FICHAS

Trabalho de conclusão de curso submetido à comissão examinadora identificada abaixo, como requisito para obtenção do grau de Bacharel em Línguas Estrangeiras Aplicadas ao Multilinguismo e à Sociedade da Informação (LEA-MSI).

---

Profa. Dra Elisa Duarte Teixeira  
Orientadora

---

Prof. Dr. Thiago Blanch Pires  
(1º. Membro da banca)

---

Profa. Ma. Fernanda de Deus Garcia  
(2º. Membro da banca)

## **AGRADECIMENTOS**

Primeiro de tudo gostaria de agradecer aos meus pais, por sempre me incentivarem a estudar e me apoiar mesmo quando eles mesmos não entendiam o que eu precisava ou estava passando, sem ter as bases que me deram eu não teria chegado tão longe. Mãe, obrigado por ter fé em mim e me incentivar todos os dias, mesmo que você mesmo não perceba isso, obrigado também por, junto do meu irmão estar comigo quando estive doente e ambos me ajudaram e dando força pela porta do meu quarto enquanto eu precisava continuar estudando e trabalhando, sem a ajuda de vocês isso tudo não seria possível.

Aos meus amigos e colegas do Lea 12, Anna, Jeferson, Anahy, Renan, Augusto e Cris. Passamos por muitas coisas nesses anos, nos momentos de diversão e amizade, que ficarão para sempre comigo.

À Anna por ter me ajudado de diversas formas na graduação, inclusive com esse trabalho, sou muito grato por tudo e feliz pela sua amizade.

Aos colegas do centro acadêmico, Aline, Sofia, Cinthia, Priscilla, Jeferson, Ian e Kathy, vivemos muitas aventuras, trabalhamos e nos divertimos muito. Vocês me ajudaram em um momento em que eu precisava de apoio e ânimo para seguir minha graduação, hoje me sinto muito mais seguro por tudo que aprendemos juntos nesse período.

Aos amigos Ian, Luiza, Giovanna, Henrique, Mateus, Helder, Carolina e Luiz Guilherme muito obrigado pelo apoio nesse ano maluco que foi 2020. Vocês estiveram juntos comigo me dando forças quando eu achava que não conseguiria, sou grato pela amizade de vocês.

Minha gratidão à minha orientadora, Elisa Teixeira, por toda a paciência comigo e por sempre me motivar em nossas reuniões e me encorajar a seguir em frente. Aos professores do LEA que contribuíram com o meu conhecimento e amadurecimento.

Aos meus colegas e professores dos grupos de pesquisa e extensão MOBILANG, em especial a Letícia, Anna e Fernanda, por compartilharem as tarefas do Banco de Intérpretes do MOBILANG. Às professoras Sabine e Carolina, que sempre confiaram no meu trabalho e me incentivaram a ser sempre melhor, obrigado por todas as oportunidades que vocês me deram de crescer.

## RESUMO

Estima-se que hoje, no mundo, existam mais de 250 milhões de migrantes internacionais, dos quais, em 2018, cerca de 68 milhões encontravam-se em situação de deslocamento forçado. No Brasil, tivemos 82.520 solicitações de refúgio apenas no ano de 2019. Tendo em vista que a barreira linguística é um dos obstáculos que têm maior impacto sobre o acolhimento de refugiados no Brasil, o grupo de pesquisa MOBILANG (Mobilidades e Línguas em Contato) decidiu criar um material de referência para prover suporte linguístico para tradutores, intérpretes, migrantes e refugiados. O *Glossário Multilíngue sobre Migração e Refúgio* foi sistematizado a partir do *Corpus Multilíngue de Migração e Refúgio* (COMMIRE), em parceria com o grupo TermiTraDiCo (Terminologia e Tradução Direcionadas por Corpus). Sua primeira versão, desenvolvida como um Trabalho de Conclusão de Curso (FURTADO, 2019), foi disponibilizada online em 2019, com 154 entradas em português brasileiro, espanhol, francês e inglês. O presente trabalho visa ampliar o *Glossário* utilizando os pressupostos teóricos e metodológicos da Linguística de Corpus para extrair novas Unidades de Tradução Especializadas (UTES) (TEIXEIRA, 2008) do COMMIRE utilizando o programa Sketch Engine (KILGARRIFF *et al.*, 2004), bem como sistematizar e descrever o processo de levantamento de UTES e de criação de novas fichas, para que futuros pesquisadores possam dar continuidade à aplicação sistemática do *Glossário*.

**Palavras-chave:** Terminografia multilíngue. Tradução especializada. Linguística de Corpus. Imigração e Refúgio

## ABSTRACT

It is estimated that there are more than 250 million international migrants in the world today, approximately 68 million of which were in a situation of forced displacement in 2018. Brazil has had 82,520 asylum claiming applications only in 2019. The research group MOBILANG (Mobilities and Languages in Contact) assessed that the language barrier is one of the obstacles that have the greatest impact on the reception of refugees in Brazil, and decided to create a reference material to provide language support for translators, interpreters, migrants and refugees. The *Multilingual Glossary on Migration and Asylum* was compiled based on the *Multilingual Corpus on Migration and Asylum* (*Corpus Multilíngue sobre Migração e Refúgio* – COMMIRE), in partnership with the TermiTraDiCo group (Corpus-Driven Terminology and Translation). The first version, developed as an Undergrad Senior Thesis (FURTADO, 2019), was published online in 2019, with 154 entries in Brazilian Portuguese, Spanish, French and English. The aim of this work was twofold: we wanted to expand the *Glossary* using a Corpus Linguistics theoretical and methodological approach to extract new Specialized Units of Translation (Unidades de Tradução Especializadas – UTEs) (TEIXEIRA, 2008) from COMMIRE with the Sketch Engine program (KILGARRIFF *et al.*, 2004). On the other hand, we were interested in systematizing and describing the process of identifying new UTEs and filling up new records for the term base, so that future researchers can continue to expand the *Glossary* on a regular basis.

Keywords: Multilingual terminography. Specialized translation. Corpus Linguistics. Immigration and Asylum.

## RESUMEN

Se estima que hoy, en el mundo, hay más de 250 millones de migrantes internacionales, de los cuales, en 2018, alrededor de 68 millones se encontraban en situación de desplazamiento forzado. En Brasil, tuvimos 82.520 solicitudes de asilo solo en 2019. Dado que la barrera del idioma es uno de los obstáculos que más impactan en la recepción de refugiados en Brasil, el grupo de **estudio** MOBILANG (Movilidades e Idiomas en Contacto) decidió crear un material de referencia para proporcionar apoyo lingüístico a traductores, intérpretes, migrantes. y refugiados. El *Glosario Multilingüe sobre Migración y Refugio* fue sistematizado con base en el *Corpus Multilingüe de Migración y Refugio* (COMMIRE), en alianza con el grupo TermiTraDiCo (Terminología y Traducción Dirigida por Corpus). Su primera versión, desarrollada como un Tesis de Conclusión del Curso (FURTADO, 2019), estuvo disponible en línea en 2019, con 164 entradas en portugués brasileño, español, francés e inglés. Este trabajo tiene como objetivo ampliar el Glosario utilizando los supuestos teóricos y metodológicos de la Lingüística de Corpus para extraer nuevas Unidades Especializadas de Traducción (UTE) (TEIXEIRA, 2008) de COMMIRE utilizando el programa Sketch Engine (KILGARRIFF *et al.*, 2004), así como sistematizar y describir el proceso de levantamiento de UTE y creación de nuevos registros, para que futuros investigadores puedan continuar la expansión sistemática del *Glosario*.

**Palabras clave:** Terminografía multilingüe. Traducción especializada. Lingüística de Corpus. Inmigración y asilo.

## RÉSUMÉ

On estime qu'il y a aujourd'hui plus de 250 millions de migrants internationaux dans le monde, dont environ 68 millions se trouvaient en situation de déplacement forcé en 2018. Le Brésil a enregistré 69.098 demandes d'asile seulement en 2019. Le groupe de recherche MOBILANG (Mobilités et Langues en Contact) a évalué que la barrière de la langue est l'un des obstacles qui ont le plus d'impact sur l'accueil des réfugiés au Brésil et a décidé de créer un matériel de référence pour apporter un soutien linguistique aux traducteurs, interprètes, migrants et réfugiés. Le *Glossaire Multilingue sur la Migration et l'Asile* a été élaboré sur la base du *Corpus Multilingue sur la Migration et l'Asile* (Corpus Multilíngue sobre Migração e Refúgio - COMMIRE), en partenariat avec le groupe TermiTraDiCo (Terminologie et traduction basées sur le corpus). La première version, développée en tant que Travail de Conclusion de Cours (FURTADO, 2019), a été publiée en ligne en 2019, avec 154 entrées en portugais brésilien, espagnol, français et anglais. L'objectif de ce travail était double: nous voulions élargir le *Glossaire* en utilisant une approche théorique et méthodologique de Corpus Linguistics pour extraire de nouvelles Unités spécialisées de traduction - UTEs (TEIXEIRA, 2008) de COMMIRE avec le programme Sketch Engine (KILGARRIFF et al., 2004). D'un autre côté, nous voulions systématiser et décrire le processus d'identification de nouveaux SEI et de remplissage de nouveaux enregistrements pour la base terminologique, afin que les futurs chercheurs puissent continuer à élargir le *Glossaire* de façon régulière.

**Mots-clés :** Terminographie multilingue. Traduction spécialisée. Linguistique de corpus. Immigration et asile.



## INTRODUÇÃO

Quando se fala em migrações no século XXI, pensa-se, em um primeiro momento, em grandes mobilidades humanas, turismo internacional, ou em migrações para países desenvolvidos em busca de uma vida melhor. Porém, há diversos outros motivos pelos quais a migração acontece. A Organização das Nações Unidas (ONU) estima que existam, hoje, cerca de 250 milhões de migrantes internacionais (International Organization for Migration, 2018); isto é, pessoas que vivem em países distintos dos que nasceram. Esse movimento também engloba deslocamento forçado: cerca de 68 milhões de pessoas no mundo se encontram na situação de refúgio (IMDH/ACNUR, 2019).

Embora atualmente os refugiados e solicitantes da condição de refúgio correspondam a uma pequena parcela da população total do Brasil, segundo o relatório “Refúgio em Números- 5ª edição”, produzido pelo Comitê Nacional para os Refugiados (CONARE), o Brasil tinha um acumulado, até dezembro de 2019, de 31.966 pessoas refugiadas e 239.706 solicitantes de reconhecimento da condição de refúgio.

Esse contexto serviu de estímulo para a criação do MOBILANG (Mobilidades e Línguas em Contato)<sup>1</sup>, grupo de pesquisa do Instituto de Letras da Universidade de Brasília dedicado ao estudo dos fenômenos linguísticos decorrentes da mobilidade e de contatos humanos, com projetos de extensão e várias linhas de pesquisa, uma das quais intitula-se “Terminologia da mobilidade”. Nos primeiros trabalhos desenvolvidos pelo grupo (MIRANDA, 2016; MOLINA CABRERA, 2017; MILITÃO, 2017; GARCIA, 2019), observou-se que a maior dificuldade dos solicitantes da condição de refúgio ao chegarem ao país era a barreira linguística. Por conta do deslocamento forçado, muitos não têm tempo de aprender o básico da língua portuguesa e enfrentam dificuldades na comunicação com os órgãos responsáveis por seu acolhimento e pela regularização de sua condição de refugiados.

Em face desse cenário, o MOBILANG criou um banco de intérpretes voluntários para atuar junto ao CONARE e à Defensoria Pública da União (DPU) (GARCIA 2019, p. 27). Além disso, em 2019, fez uma parceria com o Projeto TermiTraDiCo<sup>2</sup> (Terminologia e Tradução Direcionadas por Corpus) para compilar um banco de dados terminológico multilíngue que pudesse auxiliar esses solicitantes, bem como intérpretes, tradutores e funcionários do governo e de outras instituições não governamentais que trabalham com a temática refúgio em suas interações linguísticas.

Um primeiro passo para isso foi a compilação do COMMIRE (Corpus Multilíngue de Migração e Refúgio) (FURTADO; TEIXEIRA, 2019), ponto de partida para a construção do *Glossário Multilíngue Online sobre Migração e Refúgio* (doravante *Glossário*)<sup>3</sup>, cuja primeira versão foi apresentada por Furtado (2019) como Trabalho de Conclusão de Curso (TCC) de graduação em Línguas Estrangeiras Aplicadas ao Multilinguismo à Sociedade da Informação (LEA-MSI), na UnB.

---

<sup>1</sup> <http://mobilang.unb.br/>. Acesso em: 27 de out de 2020.

<sup>2</sup> <http://www.termitradico.unb.br/> Acesso em: 27 out 2020.

<sup>3</sup> <http://www.termitradico.unb.br/recursos/2-publicacoes/35-glossario-multilingue-online-sobre-migracao-e-refugio> Acesso em: 27 out 2020.

O presente trabalho, também apresentado como TCC de graduação para o referido curso, conta com o apoio da Cátedra UNESCO para o Multilinguismo (Capítulo Brasília) e teve por objetivo geral dar continuidade ao trabalho de exploração do COMMIRE, para ampliar o número de entradas do *Glossário*. Como objetivo secundário, nos propusemos a documentar o processo de extração de novas Unidades de Tradução Especializadas (que definiremos com vagar mais adiante) utilizando a ferramenta Sketch Engine (KILGARRIFF *et. al.*, 2004), para que futuros pesquisadores possam continuar a inserir novas entradas, tornando o *Glossário* cada vez mais útil ao público-alvo a que se destina.

No que se segue, apresentamos, primeiramente, informações básicas sobre a composição do COMMIRE e do *Glossário*, já que nos propomos a dar continuidade a esse trabalho, passando em seguida a apresentar as noções básicas da fundamentação teórica em que nosso trabalho se apoia, com destaque especial à abordagem direcionada por corpus empregada na compilação do *Glossário*, que envolve conceitos da Linguística de Corpus (doravante LC), da Tradução Especializada e da Terminografia multilíngue. Em seguida, descrevemos a metodologia empregada na elaboração das novas fichas e como foi feita a inserção dos novos verbetes no *Glossário*. Por fim, apresentamos e discutimos os principais resultados obtidos, sugerindo os próximos passos na continuidade do trabalho.

## 1. O ponto de partida: dados básicos sobre o COMMIRE e o *Glossário*

A coleta do material que forma o COMMIRE foi feita ao longo de três anos, tendo início em um trabalho de iniciação científica (FURTADO; GOROVITZ, 2017), que sistematizou o planejamento e a metodologia de coleta para compilar um corpus multilíngue sobre imigração e refúgio. O objetivo do trabalho era recolher materiais aos quais os refugiados, solicitantes de refúgio e imigrantes têm acesso quando chegam ao país para o qual vão imigrar. Esses materiais geralmente estão disponíveis na forma de cartilhas e revistas informativas, formulários, manuais e materiais didáticos (FURTADO 2019, p. 66). A soma dos materiais textuais coletados em cada língua ao final do projeto resultou no seguinte total de palavras distintas, também chamadas de *types* ou palavras-forma, e de palavras corridas, conhecidas também como *tokens* e palavras-ocorrência:

**Tabela 1:** Quantidade de *types* e *tokens* por subcorpus no COMMIRE.

Língua do Corpus	Types	Tokens
<b>Português brasileiro</b>	54.583	1.477.349
<b>Espanhol</b>	47.437	1.305.011
<b>Francês</b>	40.866	1.306.553
<b>Inglês</b>	57.332	2.091.482

Fonte: elaboração própria.

Após a compilação do corpus, o passo seguinte foi criar um banco de dados para abrigar as fichas de coleta de informações para compor o *Glossário*, cuja microestrutura oferece as seguintes informações:

- UTE – entrada do verbete, é uma Unidade de Tradução Especializada (conceito que explicaremos com mais vagar a seguir);
- Contexto – Exibe dois contextos em que a UTE ocorre;
- Variante – Exibe em que países a UTE foi encontrada (nos textos do corpus);
- Equivalente(s) – sugere forma(s) equivalente(s) nas outras três línguas do *Glossário*, com uso comprovado no corpus;
- Expressões relacionadas – UTEs relacionadas, seja pelo conteúdo semântico ou pela forma linguística.

A partir do banco de dados, foi criada uma interface de pesquisa, hospedada na plataforma App Engine<sup>4</sup> da Google, que dispensa o uso de servidor próprio. Ela utiliza a linguagem de programação Python<sup>5</sup> e o formato de arquivo YAML<sup>6</sup> para recuperar dados nas fichas e disponibilizá-los para o consulente. Em sua versão beta, o *Glossário* possuía 154 verbetes disponíveis para pesquisa, quando publicado online.

Para compor a macroestrutura da obra, Furtado (2019, p. 63) selecionou as seguintes informações:

- a) **Página “Sobre”** – Exibe as informações do projeto;
- b) **Página “Inicial”** – Exibe a caixa de busca dos verbetes;
- c) **Rodapé com “Contato”** – Fornece os dados de contato da equipe, caso o usuário queira solicitar mais informações;
- d) **Organização dos verbetes** – Mostra os verbetes do banco de dados em ordem alfabética.

A Figura 1 exibe a página de consulta do *Glossário*. A seguir, apresentamos um apanhado das bases teóricas em que o trabalho de Furtado (2019) e o nosso se apoiam.

---

<sup>4</sup> APP ENGINE. Disponível em: <https://cloud.google.com/appengine/?hl=pt-BR>. Acesso em: 25 nov 2020.

<sup>5</sup> PYTHON. Disponível em: <https://www.python.org/>. Acesso em: 25 nov 2020.

<sup>6</sup> YAML. Disponível em: <https://yaml.org/>. Acesso em: 25 nov 2020.

**Figura 1:** Página de consulta do *Glossário*.



Fonte: elaboração própria.

## 2. FUNDAMENTAÇÃO TEÓRICA

O presente trabalho, conforme mencionamos, dá continuidade a um trabalho desenvolvido por Furtado (2019), que teve por objetivo extrair Unidades de Tradução Especializadas (UTES) do COMMIRE (FURTADO; TEIXEIRA, 2019) utilizando o programa Sketch Engine e a abordagem teórico-metodológica da Linguística de Corpus. Isso significa que as entradas do *Glossário* consistem das combinações de palavras mais recorrentes no referido corpus, identificadas partindo-se de uma lista de palavras-chave em cada língua, conforme explicamos nesta seção. Antes de discorrer sobre o levantamento de novas UTES e a elaboração das respectivas fichas para o banco de dados, vamos resgatar alguns dos conceitos essenciais que embasaram o trabalho original de Furtado, já que se aplicam também ao nosso.

### 2.1 Linguística de Corpus na elaboração de glossários para tradutores

A Linguística de Corpus, segundo Berber Sardinha (2004, p. 3): “ocupa-se da coleta e exploração de corpora, ou conjuntos de dados linguísticos textuais coletados criteriosamente, com o propósito de servirem para a pesquisa de uma língua ou variedade linguística”. Este mesmo autor afirma que a LC vê a língua como um sistema probabilístico, se ocupando da observação de fenômenos linguísticos em grandes quantidades de texto, chamados de corpora (corpus, no singular), a fim de produzir generalizações sobre a língua (BERBER SARDINHA 2004, p. 30).

Os corpora podem ser classificados de acordo com diversos critérios (BERBER SARDINHA, 2004, p. 20). Neste trabalho, focalizaremos nos corpora comparáveis e multilíngues. Corpora multilíngues são aqueles que permitem a observação de textos em duas ou mais línguas. Podem ser constituídos de textos de partida e sua(s) respectiva(s) tradução(ões), alinhados ou não no nível da sentença – quando são chamados de corpora paralelos; ou de textos com características semelhantes, mas originalmente escritos em duas ou mais línguas – no nosso caso português, espanhol, francês e inglês – quando recebem o nome de corpora comparáveis (TAGNIN 2010, p. 358). Os textos que formam um corpus devem ser, na medida do possível, representativos da área à qual pertencem – em termos simples, para que tenha representatividade um corpus deve ser o maior possível, tendo um vasto número de palavras, gêneros, registros ou tipos textuais (BERBER SARDINHA, 2004, p. 25).

Tagnin e Teixeira (2012) discorrem sobre o uso de corpora e da Linguística de Corpus como abordagem na construção de obras de referência multilíngues voltadas para tradutores. As autoras reafirmam a necessidade de se trabalhar com uma variedade de materiais de referência confiáveis para que uma tradução seja de boa qualidade, particularmente no caso de textos especializados. Esses materiais devem ser capazes de auxiliar o tradutor em sua tomada de decisões durante a produção textual na língua de chegada, e por isso devem conter informações sobre a padronização da língua na área de especialidade em questão; isto é, um compilado de colocações, fraseologias e os contexto em que se inserem. Para produzir o material como este, é essencial partir de corpora de textos autênticos e produzidos naturalmente por especialistas da área.

Segundo essas autoras, para produzir um glossário direcionado por corpus (*corpus-driven*, em inglês) é importante que este seja representativo – isto é, composto de textos autênticos de um mesmo domínio, área de especialização e/ou gênero(s) textual(is) (TAGNIN; TEIXEIRA 2012, p. 56) e em quantidade suficiente para representar minimamente a área em questão. De acordo com Tagnin (2010, p. 360), espera-se que um corpus seja uma amostra representativa da língua ou variedade linguística que se pretende estudar, e cabe ao pesquisador estabelecer critérios que possam garantir essa representatividade.

No que se refere ao COMMIRE, os textos escolhidos, em sua grande maioria, foram originalmente escritos em cada uma das línguas dos subcorpora, mas há também alguns textos traduzidos, alinhados com suas respectivas traduções, coletados de fontes confiáveis - Governo, Organizações Internacionais (OIs) ou ONGs (Organizações não-Governamentais)

que recebem e/ou trabalham com esse público – e cuja procedência está bem descrita e sinalizada (FURTADO 2019, p. 41).

Na seção a seguir, introduzimos as ferramentas de LC utilizadas na extração de dados linguísticos para a criação do *Glossário*. Mas antes, apresentamos o conceito de Unidades de Tradução Especializada e sua relação com a abordagem direcionada por corpus, centrais para este trabalho.

## **2.2 Unidades de Tradução Especializada: a abordagem direcionada por corpus**

Os estudos em Linguística de Corpus podem ser divididos em dois tipos, quando se leva em conta o papel que esta ocupa na abordagem teórico-metodológica do objeto de estudo: “baseados em corpus” ou “direcionados por corpus” (McENERY; HARDIE 2012, p. 6; TAGNIN; TEIXEIRA 2012, p. 54).

A pesquisa baseada em corpus usa o corpus essencialmente para confirmar ou refutar hipóteses ou teorias linguísticas pré-definidas. Nesta abordagem, o corpus costuma ser usado como fonte de exemplos e/ou de equivalentes para termos selecionados (TAGNIN; TEIXEIRA 2012, p. 54-56; McENERY; HARDIE 2012, p. 5-6).

Em contrapartida, a pesquisa direcionada por corpus parte da observação de uma grande quantidade de textos para identificar fenômenos como a recorrência e a co-ocorrência de palavras (TAGNIN; TEIXEIRA 2012, p. 54). Aplicada à construção de obras de referência, isso significa que o conjunto de unidades que vão compor a lista de entradas da obra será fornecida pelo próprio corpus; isto é, o corpus não é usado somente como fonte de exemplos, mas como ponto de partida para identificar regularidades nos fenômenos linguísticos que caracterizam os textos de uma dada área especializada. Neste caso, a Linguística de Corpus deixa de ser uma metodologia apenas e passa a ser uma abordagem, uma forma de ver e entender a língua e seu funcionamento.

Neste trabalho, vamos dar continuidade, de uma perspectiva multilíngue, à extração das unidades que tipificam o uso da língua ou variedade linguística, a que Teixeira (2008, p. 20) deu o nome de Unidades de Tradução Especializadas (UTES), em textos recorrentes na área de imigração e refúgio, utilizando a Linguística de Corpus como abordagem. Nosso objetivo principal, como já dissemos, é ampliar o número de entradas do *Glossário* (FURTADO, 2019), uma obra de referência direcionada por corpus e voltada para a tradução.

Na próxima seção, apresentaremos as principais ferramentas da Linguística de Corpus utilizadas para a realização deste trabalho.

## **2.3 Principais ferramentas usadas nos estudos de Linguística de Corpus**

Os programas de análise linguística disponíveis hoje e usados na Linguística de Corpus, apesar de suas particularidades, costumam ter algumas ferramentas essenciais em comum:

- Listadores de palavras (*word list*): listam as palavras de um corpus de acordo com a opção selecionada (frequência, alfabética);
- Listadores de palavras-chave (*keyword list*): extraem palavras-chave, obtidas pela comparação da lista de palavras do corpus de estudo com a lista de palavras de um corpus de referência, geralmente de língua geral;
- Concordanciadores (*concordance*): Ferramenta usada para mostrar exemplos de uma expressão de busca (também chamada de nóculo) em contexto, permite a reordenação das linhas de resultado pelas palavras imediatamente à direita e à esquerda, para facilitar a identificação de padrões;

O programa escolhido para explorar o COMMIRE foi o Sketch Engine<sup>7</sup>, por vários motivos: por sua praticidade em lidar com corpora multilíngues de tamanho grande; por oferecer outros corpora multilíngues para comparação, na extração de palavras-chave; por permitir o uso de textos em *.pdf*; e também por suas funcionalidades exclusivas, como a ferramenta *Word Sketch*, que apresentamos a seguir. Durante a coleta inicial, no trabalho de Furtado (2019) e no presente artigo, utilizamos a versão paga do programa.

## 2.4 A ferramenta *Word Sketch* do Sketch Engine

Durante a rotina de levantamento e exploração de UTEs, uma ferramenta exclusiva do Sketch Engine foi imprescindível para o desenvolvimento do trabalho: *Word Sketch*. Ela processa os colocados – palavras que vêm imediatamente à esquerda e à direita da palavra pesquisada – utilizando uma versão etiquetada do corpus (em que cada palavra é marcada, automaticamente, com sua categoria gramatical), o que permite a identificação de padrões de recorrência léxico-gramaticais.

**Figura 2:** Captura parcial da tela de resultado do Word Sketch para “*asile*”.

↔ ⋮ 🔍 ✕	↔ ⋮ 🔍 ✕	↔ ⋮ 🔍 ✕	↔ ⋮ 🔍 ✕	↔ ⋮ 🔍 ✕
verbs with "asile" as object	verbs with "asile" as subject	modifiers of "asile"	adjective predicates of "asile"	"asile" is a ...
<b>demander</b> ... demander l' asile	<b>devoir</b> ... qui demande l' asile doit	<b>mineur</b> ... requérants d' asile mineurs non accompagnés	<b>recevable</b> ... votre demande d' asile est recevable	<b>personne</b> ... Un demandeur d' asile est une personne ayant fui son
<b>accompagner</b> ... requérants d' asile mineurs non accompagnés	<b>pouvoir</b> ... asile peut	<b>juste</b> ... des procédures d' asile justes et efficaces	<b>important</b> ... la procédure d' asile sont particulièrement importantes . En effet	<b>association</b> ... France terre d' asile est une association de promotion des
<b>débouter</b> ... requérants d' asile déboutés		<b>national</b> ... un système d' asile national	<b>susceptible</b> ... demande d' asile est susceptible de relever de	
<b>obtenir</b> ... obtenir l' asile		<b>efficace</b> ... des procédures d' asile justes et efficaces		
<b>déposer</b> ... demandes d' asile déposées		<b>temporaire</b> ... de l' asile temporaire		
<b>chercher</b> ... droit de chercher asile et de bénéficier		<b>relatif</b> ... les demandes d' asile relatives aux mutilations génitales		
<b>accorder</b> ... accorder l' asile		<b>équitable</b> ... politique d' asile équitable		
<b>solliciter</b> ...				

Fonte: FURTADO 2019, p. 52.

<sup>7</sup> Disponível em: <https://www.sketchengine.eu/#blue>. Acesso em 10 nov. 2020.



O recurso foi de grande ajuda para o levantamento das UTEs, visto que, ao lidar com corpora de tamanho médio-grande, é quase impossível analisar manualmente todas as linhas de concordância para algumas palavras-chave, por sua alta ocorrência. Com o *Word Sketch*, os colocados são separados por relações gramaticais, de acordo com sua frequência e respeitando medidas estatísticas de co-ocorrência. Também permite o fácil acesso a outras ferramentas do Sketch Engine, por meio de links em cada ocorrência. Em suma, o *Word Sketch* pode ser usado como uma espécie de sumário do comportamento colocacional da palavra ou expressão de busca.

Na próxima seção, falaremos sobre as etapas previstas na criação de uma ficha, de acordo com Furtado (2019) para, em seguida, detalhar os passos do processo de criação de novas fichas empregado no presente trabalho.

### 3. METODOLOGIA

Furtado (2019, p. 49-56) divide a rotina de coleta de dados e preenchimento das fichas em três partes:

- a) Coleta de palavras-chave – utilizando corpora de referência de língua geral disponíveis no Sketch Engine, são geradas e coletadas as palavras-chave a serem utilizadas para identificação de UTEs e equivalentes *prima facie*. Cada uma torna-se uma “ficha-mãe”.
- b) Coleta de UTEs com a ferramenta Word Sketch – uma vez que os resultados de pesquisa para cada palavra-chave são obtidos, é possível exportá-los em diversos formatos, como o *.xls* do Excel. Neste arquivo, observa-se as combinações recorrentes e estatisticamente significativas da palavra pesquisada e, verificado a ocorrência mínima em três arquivos distintos, gera-se uma “ficha-filha” de UTE.
- c) Extração de dados com a ferramenta *Concordance* – as UTEs de cada “ficha-filha” são pesquisadas em seu contexto de uso. As informações obtidas por meio do *Concordance* permitem preencher manualmente os campos para contextos de uso, observações, frequência e variantes.

No item 3.2.1, detalhamos e exemplificamos o processo completo de criação de novas fichas. Mas antes, é preciso falar dos mecanismos de controle de fichas do *Glossário*.

#### 3.1 Controle de Arquivos: o ponto de partida

De forma a concentrar todos os dados pertinentes ao COMMIRE e ao *Glossário*, foi criada uma planilha em formato *.xlsx* (Microsoft Excel) para abrigar todos os dados pertinentes a sua administração e controle. Em suas abas, encontramos as seguintes informações:



- a) Abas de Controle do Material que compõe os subcorpora de português, espanhol, francês e inglês, contendo o nome do arquivo, *tokens*, tipologia textual do material e país de origem;
- b) Aba de Controle de Fichas, onde todas as entradas do *Glossário* aparecem lado a lado, com seus equivalentes nas quatro línguas (quando já pesquisados), junto com sua frequência;
- c) Aba de Fichas, onde todas as fichas-mãe e fichas-filha estão disponibilizadas em ordem numérica de ID (no banco de dados), contendo os campos “entrada” e “língua”;
- d) Aba *Keyword Match*, utilizada por Furtado no desenvolvimento da versão beta do *Glossário* (FURTADO 2019, p. 51).

Havendo apresentado essas ferramentas, descrevemos o processo de criação de novas fichas a seguir.

### 3.2 Os modelos de ficha

Pensando nas necessidades do público-alvo, Furtado (2019, p. 46) definiu dois modelos de ficha padrão para o banco de dados do *Glossário*, fichas-mãe e fichas-filha.

As fichas-mãe contêm dados relativos à palavra-chave extraída, como **p.chave** (a própria palavra-chave, obtida com a ferramenta *Keywords*), **freq.t** (que significa sua frequência total no corpus, também obtida pela *Keywords*), **doc.t** (quantidade total de arquivos nos quais a palavra-chave ocorre) e, por fim, o campo **id** (que numera a ficha para controle interno do banco de dados).

**Figura 3:** Captura de tela da ficha-mãe para a palavra-chave “refugiados”.

```

pchave: refugiados
freq.t: 1.582
doc.t: 93
UTEs:
- protección de los refugiados:
  freq: 305
  doc: 46
- ex refugiados:
  freq: 8
  doc: 6
- comunidades de refugiados:
  freq: 134
  doc: 20
- ley de refugiados:
  freq: 6
  doc: 5
- protección internacional de los refugiados:
  freq: 33
  doc: 18
- reasentamiento de refugiados:
  freq: 158
  doc: 21
id:
- 0172

```

Fonte: elaboração própria.

As fichas-filha são usadas para registrar todas as UTEs recorrentes que contêm uma palavra-chave. Elas possuem os campos: **UTE**, que é a UTE em si; **UTEs\_relacionadas**, que são UTEs de conteúdo ou forma semelhantes; **contexto**, para abrigar duas sentenças extraídas do corpus com o *Concordance*; campos para abrigar as UTEs equivalentes nas outras três línguas do corpus: **pt** – português do Brasil, **es** – espanhol, **fr** – francês, **en** – inglês; **obs**, para registrar observações pertinentes; **freq**, para a frequência relativa da UTE no universo de ocorrências da palavra-chave que a originou; **def**, campo de definição da UTE (que ainda não foi preenchido, em ambos os trabalhos, mas consta para inclusão posterior); **variante**, onde são listadas as siglas ISO 3166<sup>8</sup> que correspondem ao país dos arquivos nos quais a UTE ocorre; e **id**, com a numeração das fichas no Controle de Arquivos.

<sup>8</sup> ISO-3166 COUNTRY CODES. Disponível em: <https://www.iso.org/iso-3166-country-codes.html>. Acesso em 19 nov. 2020. Vide FURTADO (2019, p. 91) para mais detalhes.

**Figura 3:** Captura de ficha-filha para a UTE “*ley de refugiados*”.

UTE:	- ley de refugiados
UTES_relacionadas:	-
contexto:	- En Uruguay, según la <b>ley de refugiados</b> y la de migración, refugiados y migrantes tienen los mismos derechos que los nacionales. - El marco legal nacional que se ocupa de los refugiados está establecido en la <b>ley de refugiados</b> 26.165, aprobada en el año 2006.
pt:	-
en:	- refugee law
fr:	-
obs:	-
freq:	- 0.1
def:	-
variante:	- UE - AR - Aml
id:	- 0173'

Fonte: elaboração própria.

### 3.2.1 Processo de criação de novas fichas

Antes de começarmos a criação de novas fichas para o *Glossário*, é necessário escolher qual será a língua de partida para a extração de dados (inglês, espanhol, francês ou português), devido à característica multilíngue do COMMIRE. Em seguida:

- Escolhemos uma palavra-chave nos resultados da ferramenta *Keyword*;
- Fazemos o levantamento das UTEs contendo esta palavra-chave, a partir do *Word Sketch*<sup>9</sup>;
- Checamos se a palavra-chave ou alguma UTE relacionada já estão na planilha de Controle de Arquivos e, em caso negativo, abrimos novas fichas;
- Selecionamos a ferramenta *Concordance* para verificar se a UTE aparece em mais de 3 arquivos;
- Buscamos dois contextos, de preferência que expliquem o significado da UTE;
- Calculamos e inserimos sua frequência relativa (vide p. 21, a seguir) no campo **freq**;
- No campo **variante**, informamos em quais países a ficha aparece, utilizando a ferramenta *Concordance*;
- Caso haja alguma observação pertinente, informamos no campo **obs**;
- Inserimos o número de controle de fichas.

<sup>9</sup> <https://www.sketchengine.eu/quick-start-guide/word-sketch-lesson/>. Acesso em 15 out 2020.

Conforme as UTEs vão sendo identificadas e as fichas-filha criadas, vamos fazendo o preenchimento da planilha correspondente no Controle de Arquivos, como mostra o excerto da Figura 4.

**Figura 4:** Captura de tela da planilha de Controle de Arquivos – UTEs contendo “asylum”.

Termo PT	freq	Termo EN	freq	Termo ES	freq	Termo FR	freq
refúgio		asylum	11303	refúgio		asile	4207
solicitação de refúgio	455	asylum application	1051	solicitud de asilo	258	demande d'asile	1.857
procedimento de refúgio	59	asylum procedure	945	procedimiento de asilo	118	procédure d'asile	423
solicitante de refúgio	588	asylum seeker	3893	solicitante de asilo	1.072	demandeur d'asile	634
pedido de refúgio	182	asylum claim	565	solicitud de asilo	-	demande d'asile	-
buscar refúgio	96	seek asylum	336	buscar asilo	63	chercher asile	18
processo de refúgio	53	asylum process	196	procedimiento de asilo	-	procédure d'asile	-
sistema de refúgio	41	asylum system	178	sistema de asilo	36	système d'asile	59
política de refúgio	11	asylum policy	123	política de asilo	17	politique d'asile	49
		asylum office	82	oficina de asilo	127		
solicitante de refúgio	-	asylum applicant	216	solicitante de asilo	-	demandeur d'asile	-

Fonte: elaboração própria

### 3.2.2 Identificando candidatos a UTE

Conforme mencionamos, uma vez selecionada a palavra-chave na ferramenta *Keyword*, é preciso abrir a planilha de Controle de Arquivos usada para registrar os termos nas quatro línguas do corpus, para ver todas as fichas já criadas para UTEs contendo uma dada palavra-chave (se houver alguma). Em seguida, passa-se à coleta de dados para as UTEs ainda não registradas, identificadas por meio do *Word Sketch*. Para exemplificar o procedimento, usaremos a colocação “*refugee claim*”, e, portanto, o inglês como língua de partida.

Após selecionar o colocado “*claim*” na ferramenta *Word Sketch*, passamos para a ferramenta *Concordance*, onde iremos ver exemplos da combinatória “*refugee claim*” nos textos que formam o subcorpus de inglês do COMMIRE. A ferramenta irá mostrar todos os contextos de uso existentes, mas é importante verificar se a combinatória ocorre em pelo menos três arquivos e, também, se é usada em três países diferentes<sup>10</sup>, o que pode ser um indicativo de que é um forte candidato a UTE, pois atesta sua representatividade. Podemos verificar essas informações clicando em uma ocorrência e verificando seu texto fonte, no *Concordance* e depois na aba do Controle de Arquivos, que registra as características dos materiais que compõem o corpus. Caso o candidato a UTE atenda a esses critérios, é aberta uma nova ficha-filha.

Ainda utilizando a ferramenta *Concordance*, localizamos dois contextos de uso para a UTE em questão. Em se tratando de unidades terminológicas, os melhores contextos são os definitórios ou explicativos, que vão ajudar o usuário do *Glossário* a entender o conceito representado, além de conferir seu uso em uma sentença. Ao registrarmos os contextos na nova ficha, a UTE deve vir em negrito – o que é feito utilizando-se a seguinte marcação: <b>

<sup>10</sup> Com exceção do português, já que o corpus contém apenas textos na variante brasileira da língua.

*refugee claim* </b>, para que, a informação estiver online e disponível no *Glossário*, o usuário possa ver a entrada em negrito (***bold***, em inglês).

No Controle de Arquivos, devemos checar a UTE na aba “controle de fichas”, para conferir se há outras fichas relacionadas e /ou UTEs equivalentes em outras línguas. Em seguida, inserimos sua ocorrência na coluna correspondente à sua língua no Controle de Arquivos. No caso de “*refugee claim*”, por exemplo, inserimos a UTE na coluna de língua inglesa e a sua frequência crua no corpus, 177, que pode ser observada no topo da página da ferramenta *Concordance*.

Algumas combinatórias podem ter formas menos frequentes, como plural, ordem invertida / indireta, ou contendo alguma palavra no meio da UTE (chamada, em inglês, de *intervening word*); p. ex.: “*comunidades de refugiados*” (Espanhol, ID 0171), aparece com mais frequência no plural que no singular; “*asylum seeking children*” (Inglês, ID 0185), que pode aparecer na ordem invertida, como e “*unaccompanied children seeking asilum*”; e “*protección de los refugiados*” (Espanhol, ID 169) que pode ocorrer como “*protección internacional de los refugiados*” (Espanhol, ID 174), em que “*internacional*” é um exemplo de *intervening word*. Nesses casos, utilizamos o campo de observações da ficha para registrar algo fora do comum. É importante lembrar que a escrita deve seguir a língua de partida da ficha. Outro exemplo de complexidade nas buscas por ocorrências no *Concordance*: a ficha “*recevoir l’asile*” (Francês, ID 177), já que “*recevoir*”, um verbo, teve que ser buscado em diferentes formas verbais (de tempo, número e pessoa), como aparecem em contexto.

Na sequência, registramos na ficha a frequência relativa (freq.) da UTE, que, diferentemente do Controle de Arquivos (onde simplesmente registramos sua frequência bruta no corpus), é preciso fazer um cálculo: a frequência da UTE é dividida pela frequência da palavra-chave no corpus. O valor é então multiplicado por cem e expresso com duas casas decimais de precisão, para indicar a porcentagem da frequência conjunta (*refugee* + *claim*) no universo de ocorrências da palavra-chave em questão (*refugee*), seguindo o cálculo de Furtado (2019, p. 49).

$$\text{freq.} = (\text{freq. UTE} \div \text{freq. palavra-chave}) * 100$$

Por exemplo, “*refugee claim*” apresentou uma frequência relativa de 1,4; isto é, das 12,328 vezes que “*refugee*” ocorreu no corpus, 177 delas foi em conjunto com a palavra “*claim*”. Em termos de porcentagem, isso significa que para cada 100 ocorrências de “*refugee*” em textos sobre imigração e refúgio, a palavra “*claim*” co-ocorre com ela 1 a 2 vezes.

Por fim, no campo variante identificamos todos os países representados nos contextos exibidos pelo *Concordance*. Isso é feito observando-se os detalhes de cada ocorrência e verificando o país de origem dos arquivos nos quais a UTE ocorre. Para “*refugee claim*”, tivemos ocorrências em textos de doze países, sendo eles: Estados Unidos (US), Grã-Bretanha (GB), Austrália (AU), União Europeia (UE), Irlanda (IE), Suécia (SE), África do Sul (ZA), Alemanha (DE), Suíça (CH), Malta (MT), França (FR) e Egito (EG). Entretanto,

conforme explicamos anteriormente, no campo variante devem ser inseridas apenas as siglas ISO, de duas letras, dos códigos dos países. Ao salvar a ficha com a extensão *.yaml*, deve-se adicionar no campo ID o seu número do controle de fichas; a nossa ficha-exemplo recebeu o número 0158’.

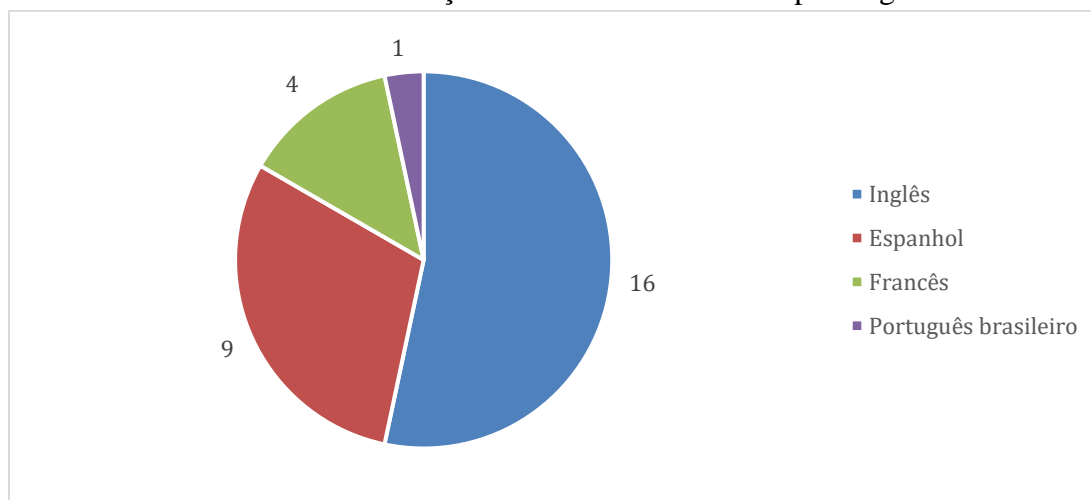
Em alguns casos, devem ser criadas novas ficha-mãe para a palavra-chave, caso ainda não existam, antes que se possa fazer o registro das ficha-filha e todas as demais fichas relacionadas. Caso a ficha já exista, basta adicionar a UTE nova, junto com sua frequência bruta no corpus, nos respectivos campos na ficha-mãe.

#### 4. RESULTADOS

Antes de passarmos à apresentação dos dados, é importante mencionar que o trabalho de Furtado teve por objetivo principal estruturar o *Glossário*, estabelecendo sua macro e microestrutura, sua interface online e o processo a ser seguido no preenchimento de novas fichas. Nosso trabalho, por sua vez, privilegiou o estabelecimento das relações de equivalência, pois entendemos que o propósito principal do *Glossário*, depois de criado, é oferecer traduções para um público que não é falante de pelo menos uma das línguas trabalhadas. Por esse motivo, em lugar de apenas compilar entradas para novas palavras-chave, priorizamos o preenchimento de fichas já criadas, especialmente nas línguas inglesa e espanhola, que complementassem as UTEs previamente compiladas nas outras línguas. No Quadro 1, as novas UTEs criadas por nós estão representadas nas células de cor verde. As células de cor azul representam a palavras-chave para a qual criamos uma fichas-mãe, “*refugiados*” (espanhol), e uma a qual fizemos acréscimos de novas UTEs, “*asile*” (francês).

Ao todo, neste trabalho, criamos 30 novas entradas para o *Glossário*, 1 ficha-mãe e 29 fichas-filha (algumas delas são novas UTEs acrescentadas à ficha-mãe da palavra-chave “*asile*”). Sua distribuição por línguas pode ser vista no Gráfico 1.

**Gráfico 1:** Distribuição das 30 UTEs coletadas por língua.



Fonte: elaboração própria.

**Quadro 1:** Fichas (e lacunas) preenchidas neste trabalho (fundo verde: UTEs novas; fundo azul: palavras-chave acrescentadas; fundo branco: UTEs existentes)

Inglês	Espanhol	Francês	Português
asylum seeking children		demandeur d'asile mineur	
Convention refugee			
former refugee	ex refugiados		
international refugee protection	protección internacional de los refugiados		
political asylum	asilo político		
quota refugee			
receive asylum	recibir asilo	recevoir l'asile	receber refúgio
refugee camps	asilo territorial		
refugee children	refugiado menor	enfants réfugiés	
refugee claim			
refugee claimants			
refugee communities	comunidades de refugiados		
refugee law	ley de refugiados		
refugee population			
refugee protection	protección de los refugiados		
refugee resettlement	reasantamiento de refugiados		
refugee status	estatus de refugiado	statut de réfugié	status de refugiado
refugee status determination			
unaccompanied children seeking asylum		requérants d'asile mineurs non accompagnés	
	condición de refugiado		condição de refugiado
	refugiados (p.chave)		
		décision d'asile négative	
		asile (p.chave)	

Fonte: elaboração própria.

Um exemplo de preenchimento de uma lacuna foi a ficha de língua espanhola “*estatus de refugiado*” (ID 0167), que completou a última lacuna do Controle de Arquivos, pois já tinha equivalentes nas outras três línguas. As UTEs “*refugee children*” (ID 0159) e “*enfants réfugiés*” (ID 0168) complementaram a ficha “refugiado menor” (ID 072), que não possuía equivalentes. Por outro lado, as UTEs “*former refugee*” (ID 0157) e “*ex refugiados*” (ID

0170) foram identificadas como equivalentes durante o processo de extração de novas entradas e, portanto, adicionadas ao *Glossário*, depois de criadas suas respectivas fichas.

Em outras UTEs de “*refugee*” (Quadro 1), temos exemplos de muitas lacunas ainda não foram preenchidas. Embora o número de equivalentes tenha sido aumentado por nós durante a fase inicial do trabalho, a maioria dessas UTEs acabaram ficando sem equivalentes pois seus correspondentes em outras línguas não preencheram algum dos requisitos para ser criada uma nova ficha – por exemplo, ter ocorrido em três arquivos diferentes representando três países diferentes.

Apenas sete das fichas novas que criamos não tiveram nenhum equivalente identificado em outras línguas – foram as primeiras que extraímos, antes de decidirmos focalizar nas fichas criadas previamente, que tinham lacunas em seus equivalentes. Ainda assim, decidimos adicioná-las ao banco de dados do *Glossário*, para que, no futuro, o trabalho possa ser continuado. Foram elas: “*quota refugee*” (ID 0155), “*convention refugee*” (ID 0156), “*refugee claim*” (ID 0158), “*refugee population*” (ID 0160), “*refugee claimants*” (ID 161), “*refugee status determination*” (ID 0165) e “*décision d’asile négative*” (ID 175).

As novas fichas serão acrescentadas ao *Glossário* assim que o presente Trabalho de Conclusão de Curso for aprovado pela banca avaliadora.

Algumas diferenças entre os corpora também foram relevantes para a extração de novas fichas. No corpus de francês, a palavra-chave “*asile*” teve frequência total de 6.790, sendo a de maior número de ocorrências; por isso, a maioria das novas fichas foi extraída dessa mesma ficha-mãe. Já em inglês e espanhol, as UTEs “*refugee*”, com a frequência total de 12.328, e “*refugiado*” com a frequência total de 9.009, respectivamente, foram as fichas-mãe com o maior número de ocorrências, portanto a criação de novas entradas nas línguas inglesa e espanhola foi mais fácil, uma vez que a semelhança entre elas facilitou a identificação de UTEs que completassem as lacunas nas duas línguas.

## 5 CONSIDERAÇÕES FINAIS

No decorrer deste projeto, apresentamos brevemente o *Corpus Multilíngue sobre Migração e Refúgio* – COMMIRE (FURTADO; TEIXEIRA, 2019) e a versão beta do *Glossário Multilíngue Online sobre Migração e Refúgio* (FURTADO, 2019) para, em seguida, descrever os passos necessários para a extração de novas UTEs e preenchimento das respectivas fichas-filha para comporem novas entradas do *Glossário*. Como aporte teórico-metodológico principal, utilizamos a Linguística de Corpus. Essa abordagem possibilita a extração de Unidades de Tradução Especializadas (TEIXEIRA, 2008) da área de refúgio e imigração em textos autênticos e representativos da área. Para a exploração do corpus e levantamento de dados para o *Glossário*, utilizamos o programa Sketch Engine (KILGARRIFF *et al.*, 2004), que possibilitou o acesso a uma grande quantidade de informações sobre a padronização por meio de suas ferramentas online, em especial *Word Sketch*.

Em termos de percalços, tivemos que manter uma atenção redobrada no preenchimento manual das fichas, visto que a probabilidade de inserção de erros é alta.



Lidamos com uma quantidade enorme de dados e estávamos em busca de detalhes importantes para a criação das fichas, o que tornou necessário revisar várias vezes os dados preenchidos, seguindo o roteiro à risca, para que nenhum dado ficasse errado ou fosse omitido. Isso demandou muito de nosso curto tempo. Outro contratempo foi a dificuldade de usar o campo “UTEs relacionadas”, visto que não conseguimos identificar o critério exato utilizado para preenchê-lo corretamente.

Algumas fichas ficaram sem equivalentes em todas as línguas pois, assim como aconteceu com Furtado, descobrimos que a extração de dados teria sido mais proveitosa se o corpus fosse ainda maior, visto que alguns equivalentes até constavam dos subcorpora, mas sua frequência não preenchia o requisito mínimo para ser criada uma nova ficha.

Outro fator que não deve ser ignorado foi o impacto da pandemia de SARS-CoV-2, que afetou o planeta neste ano de 2020, impossibilitando o pleno desenvolvimento do semestre letivo e dificultando o uso de equipamentos de informática adequados, o acesso a local próprio para estudo e, até mesmo, a saúde mental da população, de um modo geral.

Apesar das dificuldades, nosso trabalho cumpriu seu objetivo de contribuir para a ampliação do *Glossário* e, mais importante que isso, abrir caminho e oferecer as informações necessárias para que o trabalho de ampliação e atualização constante desse material seja continuado por outros participantes do projeto. Com os verbetes que compilamos e os futuros acréscimos, orientados pelo material aqui sistematizado, o *Glossário* poderá atender cada vez melhor ao seu público-alvo, cumprindo seu papel facilitador na comunicação entre migrantes e refugiados e os responsáveis por seu acolhimento no Brasil, garantindo assim seus direitos de cidadãos no país que escolheram como seu novo lar.

## REFERÊNCIAS BIBLIOGRÁFICAS

ACNUR. **Refúgio em Números**. 5ª ed. 2020. Disponível em: [https://www.justica.gov.br/seusdireitos/refugio/anexos/RESUMOEXECUTIVO\\_REFGIOE MNMEROS.pdf](https://www.justica.gov.br/seusdireitos/refugio/anexos/RESUMOEXECUTIVO_REFGIOE MNMEROS.pdf)

BERBER SARDINHA, A. P. **Linguística de Corpus**. Barueri: Manole, 2004.

FURTADO, A. B. D. e GOROVITZ, S.; Primeiros Passos Para A Compilação De Um Corpus Terminológico Sobre Situações De Mobilidade: Coleta E Análise De Duas Cartilhas Informativas Multilíngues. **23º Congresso de Iniciação Científica da UnB e 14º do DF**, 2017, Brasília.

FURTADO, A. B. D. E TEIXEIRA, E. D. Compilação de um Corpus Comparável Multilíngue de Migração e Refúgio. **25º Congresso de Iniciação Científica da Unb e 16º do DF**, 23 a 25 de setembro de 2019. Brasília, 2019.

GARCIA, F. D. **O papel do intérprete comunitário na entrevista de solicitação de refúgio**. 2019. 126 f., il. Dissertação (Mestrado em Estudos da Tradução) – Departamento de

Línguas Estrangeiras e Tradução. Universidade de Brasília, Brasília, 2019.

INTERNATIONAL ORGANIZATION FOR MIGRATION. **Global Migration Indicators**. 2018. Disponível em: [http://publications.iom.int/system/files/pdf/global\\_migration\\_indicators\\_2018.pdf](http://publications.iom.int/system/files/pdf/global_migration_indicators_2018.pdf).

KILGARRIFF, A., Pavel RYCHLÝ, Pavel SMRŽ, David TUGWELL. The Sketch Engine. **Itri-04-08**. Information Technology, 2004.

MIRANDA, Jéssica Gonçalves de Araújo. **Presas estrangeiras no Brasil: barreiras linguísticas**. 2016. 52 f., il. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2016.

MCENERY, T.e HARDIE, A. **Corpus Linguistics: Method, Theory and Practice**. Cambridge University Press, 2012.

MILITÃO, Cinthia Duarte. **O processo de pedido de refúgio e a integração acadêmica de refugiados na Universidade de Brasília**. 2017. 50 f. Trabalho de Conclusão de Curso (Bacharelado em Línguas Estrangeiras Aplicadas) — Universidade de Brasília, Brasília, 2017

MOLINA CABRERA, Marta Ingrith. **Migrações e impasses no acesso à saúde: traduzir-se é preciso**. 2017. 137 f., il. Dissertação (Mestrado em Estudos da Tradução) — Universidade de Brasília, Brasília, 2017.

SHEPHERD T. M. G.; O estatuto da Linguística de Corpus: metodologia ou área da Linguística? **Matraga vol 16**, 2009.

TAGNIN, S. E. O. Glossário de Linguística de Corpus. In: VIANA, V.; TAGNIN, S. E. O. (Org.). **Corpora no ensino de línguas estrangeiras**. São Paulo: HUB Editorial, 2010, p. 357- 361

TAGNIN, S. E. O.; TEIXEIRA, E. D. Translation-oriented, corpus-driven technical glossaries: The case of cooking terms. **Corpora vol 7**, 2012, pp. 51-67.

TEIXEIRA, Elisa Duarte. **A Lingüística de Corpus a serviço do tradutor: proposta de um dicionário de culinária voltado para a produção textual**. 2008. Tese (Doutorado em Estudos Lingüísticos e Literários em Inglês) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008. [doi:10.11606/T.8.2008.tde-16022009-141747](https://doi.org/10.11606/T.8.2008.tde-16022009-141747). Acesso em: 05 abr. 2020.